```
[1] 575.9717
```

4.4.2 De correlatiecoëfficiënt

De (Pearson) correlatiecoëfficiënt bekomen we door de covariantie te delen door de standaarddeviaties.

! De **correlatiecoëfficiënt** wordt gegeven door:

$$r_{XY} = \frac{cov_{XY}}{s_X s_Y}.$$

Het delen door de standaarddeviaties zorgt ervoor dat de correlatiecoëfficiënt tussen -1 en 1 ligt:

$$-1 \leq r_{XY} \leq 1.$$

We zullen deze eigenschap niet bewijzen.

Omdat de standaarddeviaties altijd positief zijn, heeft de correlatiecoëfficiënt hetzelfde teken als de covariantie. De correlatiecoëfficiënt voor het voorbeeld met de schoenmaten (Figuur 4.3 op pagina 115) is $r_{XY} = 1$ (perfecte positieve samenhang), voor de middelste figuur is dit $r_{XY} = -1$ (perfecte negatieve samenhang) en voor de figuur rechts geldt $r_{XY} \approx 0$ (geen samenhang).

Samengevat:

- Bij een perfecte positieve (lineaire) samenhang: $r_{XY} = 1$.
- Bij een perfecte negatieve (lineaire) samenhang: $r_{XY} = -1$.
- Indien er geen samenhang is: $r_{XY} \approx 0$.

Voor het voorbeeld rond gewicht en lengte (Figuur 4.4 op pagina 116) is $r_{XY} = 0.87$. Indien er een positieve samenhang is, maar ze is niet perfect (dus niet alle punten liggen op een rechte), zal de correlatie waarden aannemen kleiner dan 1, maar groter dan 0. Voor een negatieve samenhang die niet perfect is, zal ze negatieve waarden aannemen groter dan -1, maar kleiner dan 0.

We keren nu terug naar het voorbeeld rond Hersengrootte en Verbaal IQ. De standaarddeviaties zijn $s_X = 72.3$ voor Hersengrootte, en $s_Y = 23.6$ voor Verbaal IQ. De

covariantie is $cov_{XY} = 575.97$ (zie Tabel 4.7) zodat de correlatiecoëfficiënt gelijk is aan

$$r_{XY} = \frac{575.97}{72.3 \times 23.6} = 0.34.$$

De correlatiecoëfficiënt is positief, wat opnieuw wijst op een positieve samenhang, maar omdat ze relatief klein is, zeggen we dat de samenhang eerder 'zwak' is.

Illustratie in R

De correlatiecoëfficiënt kunnen we berekenen via `cor()`:

```
> cor(DataIQ$Hersengrootte, DataIQ$VIQ)
```

```
[1] 0.3374119
```

4.4.3 Kendall's τ

Naast de correlatiecoëfficiënt bestaan er nog verschillende andere maten van samenhang, zoals bijvoorbeeld *Kendall's* τ (*tau*). Ze wordt berekend door *concordante* en *discordante* paren te tellen.

! Een paar (x_i, y_i) en (x_j, y_j) wordt **concordant** genoemd indien:

$$\frac{y_j - y_i}{x_j - x_i} > 0.$$

! Een paar (x_i, y_i) en (x_j, y_j) wordt **discordant** genoemd indien:

$$\frac{y_j - y_i}{x_j - x_i} < 0.$$

Als voor een paar $x_i = x_j$ of $y_i = y_j$ dan is het paar niet concordant en niet discordant.

! **Kendall's** τ wordt gegeven door:

$$\tau = \frac{2(\text{aantal concordante paren} - \text{aantal discordante paren})}{n(n-1)}.$$

Merk op dat $\frac{y_j - y_i}{x_j - x_i} > 0$ wanneer ($x_i < x_j$ én $y_i < y_j$) of wanneer ($x_i > x_j$ én $y_i > y_j$). Terwijl $\frac{y_j - y_i}{x_j - x_i} < 0$ wanneer ($x_i < x_j$ én $y_i > y_j$) of wanneer ($x_i > x_j$ én $y_i < y_j$). Dus Kendall's τ maakt enkel gebruik van de volgorde van de waarden.

Analoog aan de correlatiecoëfficiënt is ook Kendall's τ begrensd door:

$$-1 \leq \tau \leq 1.$$

Verschillend van de correlatiecoëfficiënt, kan Kendall's τ ook gebruikt worden voor ordinale data.

! **Meetniveau**. Bij de berekening van Kendall's τ gebruikt men enkel de volgorde van de variabelen. Ze is bijgevolg zinnig voor ordinale, interval- en ratiovariabelen.

We illustreren deze maat aan de hand van een voorbeeld. Tabel 4.8 geeft de lengte en het gewicht voor 5 personen. We starten met alle personen paarsgewijs te vergelijken:

- Persoon 1 ($i = 1$) en persoon 2 ($j = 2$). Voor deze personen geldt $x_1 = 160$, $y_1 = 53$, $x_2 = 168$ en $y_2 = 55$. Dit paar is concordant omdat $\frac{y_2 - y_1}{x_2 - x_1} = \frac{55-53}{168-160} = \frac{2}{8} = 0.25 > 0$.
- Persoon 1 ($i = 1$) en persoon 3 ($j = 3$). Voor deze personen geldt $x_1 = 160$, $y_1 = 53$, $x_3 = 176$ en $y_3 = 52$. Dit paar is discordant omdat $\frac{y_3 - y_1}{x_3 - x_1} = \frac{52-53}{176-160} = \frac{-1}{16} < 0$.
- Analoog voor de overige paarsgewijze vergelijkingen: persoon 1 en persoon 4 (concordant), persoon 1 en persoon 5 (concordant), persoon 2 en persoon 3 (discordant), persoon 2 en persoon 4 (concordant), persoon 2 en persoon 5 (concordant), persoon 3 en persoon 4 (concordant), persoon 3 en persoon 5 (concordant), persoon 4 en persoon 5 (concordant).

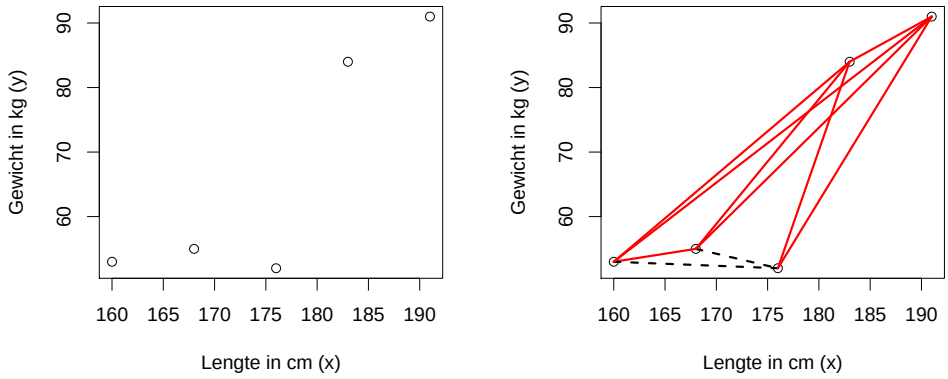
Er zijn dus 8 concordante paren en 2 discordante paren, zodat

$$\tau = \frac{2 \times (8 - 2)}{5 \times (5 - 1)} = \frac{12}{20} = 0.6.$$

De concordante en discordante paren kunnen we ook visueel voorstellen door alle punten in het spreidingsdiagram paarsgewijs te verbinden via rechten, zoals weergegeven in Figuur 4.6. De rechten met een positieve richtingscoëfficiënt (een rechte van linksonder naar rechtsboven) zijn de concordante paren (rode volle lijn op de figuur) en de rechten

Persoon (i)	Lengte in cm (x_i)	Gewicht in kg (y_i)
1	160	53
2	168	55
3	176	52
4	183	84
5	191	91

Tabel 4.8: Dataset om de berekening van Kendall's τ te illustreren.



Figuur 4.6: Links: spreidingsdiagram voor de data uit Tabel 4.8. Rechts: spreidingsdiagram met aanduiding van de concordante paren (rode volle lijn) en discordante paren (zwarte stippellijn).

met een negatieve richtingscoëfficiënt (een rechte van linksboven naar rechtsonder) zijn de discordante paren (zwarte stippellijn op de figuur).

Er bestaan verschillende formules om Kendall's τ te berekenen wanneer een waarde meerdere malen voorkomt ($x_i = x_j$ of $y_i = y_j$), maar deze worden niet besproken in de cursus.

Voor de spreidingsdiagrammen in Figuur 4.3 op pagina 115 geldt dat $\tau = 1$ voor de perfect positieve samenhang (figuur links), $\tau = -1$ voor de perfecte negatieve samenhang (figuur midden) en $\tau \approx 0$ wanneer er geen samenhang is (figuur rechts).

Voor Hersengrootte en Verbaal IQ geldt dat $\tau = 0.28$, wat opnieuw wijst op een (zwakke) positieve samenhang.

Illustratie in R

Door de optie `kendall` te gebruiken, kunnen we via `cor()` Kendall's τ berekenen:

```
> cor(DataIQ$Hersengrootte, DataIQ$VIQ, method = "kendall")
```

```
[1] 0.2839256
```

Deze berekening maakt gebruik van speciale rekenregels om om te gaan met waarden die meerdere malen voorkomen. Zoals eerder aangegeven gaan we hier niet dieper op in.

4.4.4 Lineaire en niet-lineaire verbanden

We hebben in de voorgaande paragrafen twee verschillende maten gezien voor samenhang die beiden begrensd zijn tussen -1 en 1, namelijk de correlatiecoëfficiënt en Kendall's τ. Deze laatste maat kan gebruikt worden voor variabelen van tenminste ordinaal meetniveau terwijl de correlatiecoëfficiënt enkel kan gebruikt worden voor variabelen van ten minste interval meetniveau. Er is echter nog een ander belangrijk verschil tussen beiden maten: de correlatiecoëfficiënt (en de covariantie) is een maat voor de *lineaire* samenhang tussen twee variabelen terwijl Kendall's τ een maat is voor een *monotone* samenhang. We bespreken eerst kort wat lineaire en monotone functies zijn.